Data Analysis Techniques

Steve Park
Department of Computer Science
College of William and Mary
Williamsburg, VA

A large and diverse number of computational techniques are routinely used to process and analyze remotely sensed data. These techniques include:

- Univariate Statistics;
- Multivariate Statistics;
- Principal Component Analysis;
- Pattern Recognition and Classification;
- Other Multivariate Techniques;
- Geometric Correction;
- Registration and Resampling;
- Radiometric Correction;
- Enhancement;
- Restoration;
- Fourier Analysis;
- Filtering.

Each of these techniques will be considered, in order.

Univariate Statistics

The standard way to reduce a large amount of *homogeneous* remotely sensed data to a handful of representative numbers is to apply traditional elementary univariate statistical techniques. These include:

- measures of central tendency (mean, median, mode);
- measures of dispersion (variance, standard deviation);
- measures of distribution (percentiles, histograms).

These techniques are well understood and only require a modest computational capability, particularly if they are implementated as "one-pass" algorithms. That is, for example, the variance, s^2,

(and the mean, $\overline{x}$) of n data points should be calculated in one pass as

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \overline{x}^2$$

rather than in two passes as

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Doing so minimizes memory requirements.

As an alternative, if the data is quantized as, say, 8-bit integers then a 256-bin histogram is easily computed and from it all other statistics can be calculated. For example, the mean is then

$$\overline{x} = \frac{1}{n} \sum_{l=0}^{255} l f[l]$$

where $f[l]$ is the histogram bin count corresponding to the data value l.

Multivariate Statistics

The standard way to compare several related sets of remotely sensed data (each of which is homogeneous)—for example, multi-spectral data—is to apply traditional multivariate statistical techniques. These include:

- measures of co-relation (covariance matrices, correlation matrices);
- measures of multivariate distribution (bivariate histograms).

As in the case of univariate statistics, these techniques are well understood and only require a modest computational capability, particularly if they are implemented as one-pass algorithms. Thus, for example, the covariance, c, (and the associated means) of two sets of n data points should be calculated in one pass as

$$c = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}$$

rather than in two passes as

$$c = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

As an alternative, multivariate statistics can be calculated from bivariate histograms. For example

$$\sum_{i=1}^{n} x_i y_i = \sum_{l_1=0}^{255} \sum_{l_2=0}^{255} l_1 l_2 f[l_1, l_2]$$

where $f[l_1, l_2]$ is the bivariate histogram bin count corresponding to the data pair (l_1, l_2). However, the desirability of bivariate histograms as analysis tools for multivariate remotely sensed data is problematic. For example, a complete set of bivariate histograms for a 6-channel, 8-bit multispectral data set would involve the use of 15 arrays, each 256×256.

Principal Component Analysis

Multivariate remotely sensed data frequently exhibits inter-component correlation; e.g., adjacent spectral bands are highly correlated. For this reason the *principal component transformation* (PCT) is sometimes applied to "decorrelate" the data. The result is a new (uncorrelated) set of data. If the multivariate data is k dimensional, then the PCT is implemented as a $k \times k$ matrix which is applied on a (multivariate) data point-by-data point basis. The amount of computation per data point is typically modest, provided k is not large. However, the construction of the PCT matrix involves a significant amount of computation. Three steps are involved.

- Calculate the $k \times k$ covariance matrix (or an approximation thereof) for the original multivariate data set.
- Calculate the eigenvector - eigenvalue decomposition of the covariance matrix.
- Construct the PCT matrix from the k eigenvectors of the covariance matrix.

Since the data must be processed once to calculate the PCT matrix and then a second time to apply the transform, principal component analysis is an inherently two-pass algorithm with a corresponding potentially large memory requirement. To get around this problem, it is common to essentially "guess" the covariance matrix a priori thereby avoiding one of the two passes.

If T is the PCT matrix and $\mathbf{x}$ is a k-dimensional multivariate data sample, then the transformation

$$\mathbf{x}' = T\mathbf{x}$$

defines a new set of multivariate data. It is frequently the case that the (diagonal) correlation matrix for this new data set will have only a few significant diagonal elements indicating that the effective dimension k' of the new data set is less than that of the original, k. This dimensionality reduction is an important consideration in pattern recognition and classification.

Pattern Recognition and Classification

Pattern recognition and classification is a statistical technique frequently applied to multispectral, remotely sensed image data. The concept is easily understood, however a successful implementation is frequently difficult. The primary reason for this is the difficulty of providing a sufficiently accurate statistical characterization of the classes.

The computational requirements associated with pattern recognition and classification are similar to those associated with a principal component analysis. Indeed, the PCT is frequently advocated as a preprocessing dimensionality reduction step so that classification can then be accomplished in a new smaller and "more orthogonal" space in which patterns become better separated. The point is that classification is done on a sample-by-sample basis with the amount of computation per classification determined by the complexity of the classification algorithm. There is a classic accuracy versus time tradeoff involved—more accurate classification is typically associated with a more complex algorithm. The complexity is, in turn, usually determined by the amount of statistical data (assumptions) required to characterize the patterns.

Most classification techniques are based upon a *maximum likelihood* application of *Bayes' Rule*. That is, each possible k-dimensional measurement vector $\mathbf{x}$ is assumed to have a probability distribution $\Pr\{\mathbf{x} \mid c\}$ for each of a finite number of pattern classes, c. The a priori probability of each class, $\Pr\{c\}$ is also assumed to be known. The conditional probability that $\mathbf{x}$ belong to class c is then

$$\Pr\{c \mid \mathbf{x}\} = \frac{\Pr\{c\}\Pr\{\mathbf{x} \mid c\}}{\Pr\{\mathbf{x}\}}$$

where

$$\Pr\{\mathbf{x}\} = \sum_{c} \Pr\{c\} \Pr\{\mathbf{x} \mid c\}$$

and the sum is over all possible classes. A measurement vector $\mathbf{x}$ is then classified by finding (searching for) that class c for which $\Pr\{c \mid \mathbf{x}\}$ is largest.

Other Multivariate Techniques

There are an enormous number of largely empirical techniques which have been developed for specific types of multivariate remotely sensed data. Most of these techniques attempt to answer the same basic question . . . "given two sets of data which may be separated in time, space, or wavelength to what extent are the two *different*?" It is impossible to comprehensively summarize all these techniques. However, it is at least possible to demonstrate three classifications into which many of these techniques fall.

- Multispectral—two nearly simultaneous observations of the same spatial area but in different spectral bands; e.g., vegetation studies in the visual and near IR. In an attempt to correct for atmospheric effects, many techniques involve "ratioing"—the division of one data set by the other. Because multispectral operations are typically quite simple and are applied sample by sample, the computational requirements are usually modest.

- Multitemporal—two spectrally compatible observations of the same spatial area made at significantly different times; e.g., seasonal observations of crops. The basic issue here is *change-detection* and many techniques involve "differencing", which is the subtraction of one data set from the other. The operations here are also typically quite simple and are applied sample by sample. As a result, the computational requirments are usually modest.

- Multispatial—two nearly simultaneous and spectrally compatible observations of the same area but differing in orientation, spatial resolution, or both; e.g., two similar imaging systems viewing the same spatial area but from different points in space. This geometric process of mapping one data set onto another is called *spatial registration* and is a notoriously complex computational process, particularly for image data.

Many multivariate techniques are actually "multi-multi". For example, it is frequently the case that one wants to analyze two images of approximately the same spatial area taken at different times, by different instruments located at different points in space. This is always a non-trivial process.

Geometric Correction

A major concern in the processing and analysis of remotely sensed data is to ensure that the data has the correct geometric characteristics. This is particularly true for image data. The details of the application and the instrument determine the requirements and therefore the techniques. However, the following factors are all potentially important and their contribution to a lack of geometric fidelity must be understood:

- detector geometry—defines the sample (pixel) grid in the scene;

- analog filter and detector response delays—a lag in time (perhaps frequency dependent) which may produce a potentially significant spatial shift in the data;

- sensor scan rate and slew (if applicable)—the ground track corresponding to a scanning sensor is "curved" relative to the ground track of the instrument;

- spacecraft altitude and velocity—departures from nominal can produce potentially significant geometric effects;

- spacecraft attitude (roll, pitch, and yaw)—departures from nominal can produce potentially significant geometric distortion;

- earth rotation;

- perspective geometry.

Geometric correction *in real time* will be a computational challenge. This is particularly true for image data. Recognize that the "traditional" way to do a final geometrtic correction (if high geometric fidelity is required) is on the ground with a computer-aided, interactive system heavily reliant on ground control points.

Registration and Resampling

Multispatial problems of registration and resampling arise primarily with image data whenever pixel values are required at points where they don't exist. Spatial registration is the geometric process whereby one coordinate system is mapped onto another. In effect, an image is created on a rubber sheet which is then pulled and stretched onto a new coordinate (pixel) grid. New pixel values must then be generated—this interpolation process is called resampling. Many resampling techniques are available, however computational consideration virtually always dictates that only local, efficient techniques are used. Local methods typically require a knowledge of the 16 nearest neighboring pixels (on a 4×4 grid) and their associated pixel values. This presents a significant computational challenge if the image is not in fast, random access memory.

In practice, resampling is virtually always implemented using either

- bilinear interpolation—in which case only the 4 nearest neighboring pixels are involved, but the interpolation function is not "smooth", or

- parametric cubic convolution—a smooth interpolation process involving the 16 nearest neighboring pixels.

The application (spatial resolution) determines which of these techniques should be used. If high spatial resolution is the requirement, cubic convolution is the choice.

Multispectral problems of registration and resampling can also arise. They are much less commonly discussed in the (image processing type) literature. In this case, the problem is really a restoration or inversion problem because spectral bands are frequently not sufficiently narrow enough to permit multispectral data to be analyzed as "point" (wavelength) samples. However, although the mathematics here may be involved, the implementation typically requires only a modest computational capability, unless the number of spectral bands is large.

Radiometric Correction

Another major concern in the processing and analysis of remotely sensed data is to ensure that the data is radiometrically correct. The following factors are all potentially important and their contribution to a lack of radiometric fidelity must be understood:

- detector response—the conversion of sensed radiance into voltage, typically modeled as an I/O transfer function with multiple parameters (gain, offset, etc.) whose values are estimated via extensive extensive pre-flight and in-flight calibration tests;
- analog-to-digital (A/D) conversion—the conversion always causes a quantization error; this error is negligible only if the number of bits per sample is sufficiently large enough and the "dynamic range" of the converter is properly matched to the detector response;
- atmospheric effects—the atmosphere acts as a (stochastic) filter to modulate the "signal" as it passes from the scene to the detector.

Correction for the first two factors is typically straightforward with only a modest computational effort required—unless extremely high radiometric fidelity is required; in which case, the detector response model may be quite complex (e.g., a Kalman filter). Even in this case, however, the correction is applied sample by sample and so the memory requirements are modest.

Atmospheric corrections *in real time* will be a computational challenge primarily because of the need to combine data from a variety of sensors and apply a significant amount of "engineering judgement" when these data do not all agree.

Enhancement

Enhancement is any *empirical* technique applied to data (typically image data) to increase the "information content" (visually or otherwise) of the data. This is the stuff of image processing and includes a variety of techniques which are usually applied pixel by pixel. That is, if $g(m,n) = l$ represents the value (gray level) of pixel (m,n) and if $l' = T[l]$ represents a gray level transformation (typically stored as a look-up-table) then the transformation

$$g'(m,n) = T[g(m,n)]$$

applied for all pixels (m,n) defines a new (enhanced) image, g'. The choice of $T[\cdot]$ defines the enhancement and includes the following standard techniques:

- linear contrast stretching—linearly stretch all values in the primary range of interest and clip the others;
- histogram equalization—redistribute gray levels so as to maximize the entropy of the enhanced image;
- histogram specification—redistribute gray levels in accordance with a specified (desired) histogram;
- gray level slicing—selected gray levels are unchanged; all others are clipped.

Pseudocolor enhancement is an extension of gray level enhancement in which scalar data is mapped to a color display by the pseudo (false) assignment of colors to gray levels. This is an increasingly popular technique in remote sensing (with limited scientific merit) whereby even the most benign data set can be manipulated to produce a quite dramatic appearance.

Restoration

Radiometric correction is really just a special case of the *restoration* (or inversion) problem. For example, an image of a scene is only an imperfect copy of that ideal image which would be produced if it were possible to geometrically project the scene through the optical system with no degradation. The traditional discrete 2-d *linear* model of this process is

$$g(m,n) = \sum_{m'} \sum_{n'} h(m,n;m',n')f(m',n') + e(m,n)$$

where f represents the ideal image of the scene, h represents the system impulse response (point spread function), e is an (additive) noise term, g is the actual measured image, and the summation is over all space. Both g and h are assumed to be known and it is presumed that the noise, e, can be characterized statistically. The problem then is to solve for f. Equivalently, remove the effects of the instrument (h) and the noise (e) from the data (g) to determine the object (f) which was remotely sensed.

In many applications the instrument is so well designed (h is effectively a delta function) and the noise is so small ($e \approx 0$) that the subtle difference between f and g can be ignored. When this is not the case, the restoration problem must be solved. Doing so is a mathematically and computationally challenging activity which is usually facilitated by the additional assumption that the linear process which relates f to g is also *shift-invariant*. In this case, the previous equation simplifies to a *convolution*

$$g(m,n) = \sum_{m'} \sum_{n'} h(m-m',n-n')f(m',n') + e(m,n)$$

and the problem can be analyzed and solved (at least in theory) by using Fourier methods.

Fourier Analysis

If g is a 2-d array (e.g., an image) of size $M \times N$, then the corresponding *discrete Fourier transform* (DFT) of g is the $M \times N$ array $\hat{g}$ defined by

$$\hat{g}(\mu,\nu) = \frac{1}{MN} \sum_{m} \sum_{n} g(m,n) \exp\left(-2\pi i \left(\frac{m\mu}{M} + \frac{n\nu}{N}\right)\right).$$

The definition in 1-d is analogous and the continuous form of this transform involves little more than replacing $\sum$'s with $\int$'s.

There are several good reasons why the Fourier transform enjoys such popularity:

- the convolution theorem—convolution in the spatial (m,n) domain is equivalent to multiplication in the frequency (μ,ν) domain;
- many remote sensing instruments are built to specifications (resolution, response, etc.) which are formulated in the frequency domain;
- the DFT lends itself to the analysis of periodic data;

- there is a compelling "duality" between space (or time) and frequency; e.g., the array g can be recovered from the $\hat{g}$ array by a second (inverse) Fourier transform

$$g(m,n) = \sum_{\mu} \sum_{\nu} \hat{g}(\mu,\nu) \exp\left(2\pi i \left(\frac{m\mu}{M} + \frac{n\nu}{N}\right)\right);$$

- for most values of M and N there is a *fast* implementation of the DFT—the fabled *fast Fourier transform* (FFT).

All of this must be balanced against the observation that, for large values of M and N the FFT is a computational resource hog—memory requirements are proportional to MN and the operations count is proportional to $MN \log_2(MN)$.

Filtering

Filtering is the process of modifying sample values using a weighted linear combination of sample values in a neighborhood of the value in question. For example, if g is a 2-d array of data and h is a 2-d array of weights then the (convolution) equation

$$g'(m,n) = \sum_{m'} \sum_{n'} h(m-m', n-n') g(m', n')$$

defines a new, filtered, array g'. In 1-d and 2-d, this is the stuff of signal and image processing.

Filtering is typically used for:

- noise suppression—a "low-pass" filter designed to suppress high-frequency noise;
- high frequency enhancement—a "high-pass" filter designed to empirically boost high frequencies thought to be previously suppressed during data acquisition;
- high frequency restoration—typically, some form of a Wiener filter derived based upon a convolutional model of the data acquisition process.

In most cases, filters are designed (or derived) in the frequency domain and then transformed to the spatial domain. The design of a filter is always a challenging process. If (and perhaps, only if) it is done right, then the spatial domain implementation in 1-d requires only a modest computational capability. In 2-d, however, there is likely to be a data-management problem associated with maintaining a data list for the nearest neighbors of the point being filtered.

References

For those interested in acquiring a better background in data analysis techniques as they relate to remote sensing, I recommend beginning with the

- *Manual Of Remote Sensing*, 2nd edition, (ISBN 0-937294-41-1) chapters 17, 18, and 21

and the many references contained therein. Although this manual is now 6 years old, much of the material in these three chapters remain state-of-the-art. Chapter 17 (edited by Fred Billingsley) is particularly recommended. However, be aware that the computational view in each of these chapters is the traditional one—do all the processing on the ground, take as much time as necessary, and use all the horsepower you can muster.